# The Development and Validation of a Neural Model of Affective States

Katherine Lorraine McCurry

Thesis submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Master of Science

In

Psychology

Brooks King-Casas, Committee Chair

Pearl H. Chiu

Stephen M. LaConte

Susan W. White

September 23, 2015

Roanoke, VA

Keywords: fMRI, emotion, support vector machine, machine learning, neurofeedback

The Development and Validation of a Neural Model of Affective States

Katherine Lorraine McCurry

## Abstract

Emotion dysregulation plays a central role in psychopathology (B. Bradley et al., 2011) and has been linked to aberrant activation of neural circuitry involved in emotion regulation (Beauregard, Paquette, & Lévesque, 2006; Etkin & Schatzberg, 2011). In recent years, technological advances in neuroimaging methods coupled with developments in the field of machine learning have allowed for the non-invasive measurement and prediction of brain states in real-time, which can be used to provide feedback to facilitate regulation of brain states (LaConte, 2011). Real-time functional magnetic resonance imaging (rt-fMRI)-guided neurofeedback, has promise as a novel therapeutic method in which individuals are provided with tailored feedback to improve regulation of emotional responses (Stoeckel et al., 2014). However, effective use of this technology for such purposes likely entails the development of (a) a normative model of emotion processing to provide feedback for individuals with emotion processing difficulties; and (b) best practices concerning how these types of group models are designed and translated for use in a rt-fMRI environment (Ruiz, Buyukturkoglu, Rana, Birbaumer, & Sitaram, 2014).

To this end, the present study utilized fMRI data from a standard emotion elicitation paradigm to examine the impact of several design decisions made during the development of a whole-brain model of affective processing. Using support vector machine (SVM) learning, we developed a group model that reliably classified brain states associated with passive viewing of positive, negative, and neutral images. After validating the group whole-brain model, we adapted this model for use in an rt-fMRI experiment, and using a second imaging dataset along with our group model, we simulated rt-fMRI predictions and tested options for providing feedback.

*Keywords*: fMRI, emotion, support vector machine, machine learning, neurofeedback

## Table of Contents

## List of Tables

## List of Figures

## Introduction

Emotion processing is essential to adaptive functioning, and emotion dysregulation is believed to be a transdiagnostic risk factor for psychopathology (Joormann & Goodman, 2014). Emotion dysregulation has also been linked to increased severity of psychiatric symptoms (B. Bradley et al., 2011). For example, research on posttraumatic stress disorder (PTSD) has shown emotion dysregulation to be significant predictor of PTSD status in trauma-exposed individuals (Weiss, Tull, Anestis, & Gratz, 2013), to be related to the severity of PTSD symptoms (Tull, Barrett, McMillan, & Roemer, 2007), to moderate the relation between PTSD symptom severity and substance use (Tull, Bardeen, DiLillo, Messman-Moore, & Gratz, 2014), and to be associated with frequency of self-harm behaviors (Dixon-Gordon, Tull, & Gratz, 2014).

Given the significant role of emotion processing in daily functioning and the adverse consequences that result from deficits in emotion regulation, researchers have sought to understand the neural underpinnings of both healthy and dysfunctional emotion processing and regulation. Early work based on animal models of emotion and experiments with patients with brain lesions implicated the limbic system as the seat of emotion processing (Dalgleish, Dunn, & Mobbs, 2009). However, since the advent of modern neuroimaging technology, substantial research on the neural correlates of emotion processing in healthy human subjects has shown emotion processing involves not only limbic regions including the amygdala, thalamus, hippocampus, and hypothalamus, but also widespread activity in cortical regions including the medial prefrontal cortex (mPFC), orbitofrontal cortex (OFC), and insular cortex (Phan, Wager, Taylor, & Liberzon, 2002). Likewise, neuroimaging research in psychopathology has found emotion dysregulation related to psychiatric disorders to be associated with abnormal patterns of activity in many of the same regions implicated in emotion processing (Beauregard et al., 2006;

Cisler et al., 2013; Etkin & Schatzberg, 2011; Taylor & Liberzon, 2007). For instance, results suggest that during the down-regulation of negative emotions, individuals with depression show increased activation of the right amygdala, right insula, right anterior temporal pole, and right dorsal anterior cingulate cortex (dACC) compared to control participants (Beauregard et al., 2006).

The knowledge of neural deficits associated with emotion dysregulation provide a likely target for a biologically informed treatment such as neurofeedback, a technique that aims to facilitate learning by providing an individual with real-time information about his/her brain state. Real-time fMRI-guided neurofeedback, in contrast to neurofeedback based on EEG, has the ability to provide precise localization and can measure activity in deep structures non-invasively (Linden, 2014). In addition, as emotion is associated with global activity across multiple networks in the brain, multivariate analytic approaches may provide better characterization of areas of significance than univariate approaches do. Real-time fMRI-guided neurofeedback that employs multivariate analyses to classify brain states has promise as a novel therapeutic method in which individuals are provided with tailored feedback to improve regulation of emotional responses (Stoeckel et al., 2014).

To this end, the present study aimed to use fMRI data from a standard emotion elicitation paradigm to examine the impact of several design decisions made during the development of a whole-brain group model of affective processing for eventual use in rt-fMRI studies of emotion regulation. Prior to detailing our study, we have provided background regarding key components of our study: emotion and its neural correlates, machine learning approaches to brain state classification, and real-time fMRI.

## Emotion

### Psychological perspectives.

***Scientific framework***. For any scientific approach to the investigation of emotion, an important consideration is what can be measured with regard to emotion or what are the data of emotion. In the field of psychology, targets for studying emotion usually fit into one or more of the following four categories: (a) somatic/bodily responses; (b) behavioral responses; (c) cognitions; and (d) feelings (Frijda, 2010; Lang, 2010). Not surprisingly, these categories are also the components most often included in psychological definitions of emotion. However, researchers define these components in different manners (Moors, 2009). For example, some researchers subdivide the second target (behavioral responses) into two related components: (a) a motivational piece (e.g., action readiness or tendencies towards action) and (b) a motor piece (e.g., actions, verbal and non-verbal expressions) (Moors, 2009). In addition, some researchers have used a broad definition of the third target (cognitions) while others have defined it more narrowly (Moors, 2009). Also, the fourth target (feelings or the conscious 'experience' of an emotion) has been considered subjective by many, and there is debate about whether it should be considered a measurable target of scientific investigation (Lang, 2010). Finally, in addition to being a vehicle for the experimental manipulation of emotion, the properties of the stimulus that elicits the emotion response have also been a target of study (Britton, Taylor, Sudheimer, & Liberzon, 2006; Schlochtermeier et al., 2013; Wright et al., 2003)**.**

Moreover, there are several levels at which these targets might be measured. Daniel Dennett (as cited in Frijda, 2010 pg.78) identified the following three descriptive levels at which inquiry could occur: (a) an intentional level; (b) a functional level; and (c) a neural level. A

target, such as cognition, might be studied at more than one of these three levels, and there also may be smaller levels of description within the currently identified levels (Frijda, 2010).

**Historical theories.** An overview of the history of psychological theories of emotions can be examined using the targets of study, the levels of description, and the distinction between dimensional/constructed or discrete/basic theories as organizing principles. Much of the early psychological research on emotion by individuals such as Wilhelm Wundt, William James, and Carl Lange was informed by the work of their contemporaries in philosophy and biology. As such, these early scientists' work was in response to controversies not only within their field but also within the broader study of emotion.

***James-Lange theory.*** In 1884, William James, who is widely regarded as the father of American psychology, wrote a paper that challenged the commonly held definition of emotion. The accepted theory of emotion at that time was that a mental perception of a stimulus resulted in a feeling or emotion state and then that feeling/emotion state led to bodily reactions. James's theory countered this by postulating that mental perception of a stimulus was followed by bodily reactions, and it was the interpretation of those bodily sensations that resulted in the emotional feeling (James, 1884). Around the same time, Carl Lange, a Danish psychologist, also proposed a similar theory, stating, "We owe all the emotional side of our mental life, our joys and sorrows, our happy and unhappy hours, to our vasomotor system" (as cited in Cannon, 1927 pg. 107). This theory, now known as the James-Lange theory, shifted the focus from the subjective feeling of emotion towards the physiological responses to environmental stimuli as representative of emotion itself (Dalgleish, 2009). In doing so, it attempted to move the level of description from the intentional to the functional. It also was in line with a dimensional conceptualization of emotion (L. F. Barrett & Wager, 2006). As the first complete psychological theory of emotion,

this theory has continued to be influential despite evidence that seems to disprove at least portions of it.

***Cannon-Bard theory.*** In 1927, Walter Cannon wrote a paper challenging the James-Lange theory. In this paper, he discussed research in which scientists surgically removed the sympathetic portion of the autonomic nervous system of cats and demonstrated that even without this part of their nervous system, the animals still engaged in emotional behavior (Cannon, 1927). Canon (1927) used this result to argue that emotion could not require the perception of physiological changes as the James-Lange theory proposed because the animals exhibited emotional behavior despite an inability to perceive visceral changes due to an absent sympathetic nervous system. Instead, Cannon proposed an alternate theory, which states that "the peculiar quality of the emotion" is felt in addition to bodily responses when the thalamic processes are stimulated (1927 pg. 120). He hypothesized that this could occur without direct involvement of the cortex and suggested that the cortex was instead responsible for modulation of initial reactions as well as choosing appropriate actions in response (Cannon, 1927; Dalgleish, 2010). This theory, which is now known as the Cannon-Bard theory, provided the first theoretical brain-based account of emotion (Dalgleish, 2010). In doing so, the theory engaged both the functional and neural levels of description.

***Schachter-Singer two-factor theory.*** In 1962, Stanley Schachter and Jerome Singer published a paper that proposed the addition of cognitive factors to the theory of emotion processing. In this paper, they suggested that physiological arousal is a necessary but not sufficient condition for emotion (Schachter & Singer, 1962). They theorized that when individuals experience physiological arousal in absence of a readily available explanation, they label the arousal state based on their cognitions at the time (Schachter & Singer, 1962). This

theory was in support of the dimensional framework of emotion and engaged both the functional and neural levels of description.

To investigate this theory, the researchers performed an experiment in which participants were administered a shot containing epinephrine or a placebo. The dosage of epinephrine administered typically resulted in a quick increase in blood pressure, heart rate and respiration. Participants were divided into groups with one group being informed of the harmless side effects and one group not being informed of any side effects; there was also an additional control group that was told the drug caused different, inaccurate side effects (itching feet and head). The researchers hypothesized that individuals who were informed about the actual side effects would have an available interpretation for the change in their physiological state and thus, would not be likely to look to other cognitive interpretations such as emotions. However, they believed the individuals in the uninformed group, having no available explanation for the symptoms, would look to available cognitions to interpret and label the experience of the side effects. The researchers sought to manipulate the cognitions of the participants by placing them in a room with a confederate who had been instructed to either act euphoric or angry. Overall, the results of the experiment supported the researchers' hypothesis that those individuals without an explanation for their physiological state were susceptible to the confederate manipulation into either the euphoria or anger state while those who had been informed of the correct side effects did not show the same susceptibility (Schachter & Singer, 1962). However, not all of the data was in line their original hypotheses and the researchers, themselves, acknowledged that the experimental manipulations did not have clear real-world translations (Dalgleish, 2009). Despite these problems, the resulting theory, referred to interchangeably as the Schachter-Singer or the two-factor theory of emotion, was accepted in the field (Dalgleish, 2009).

***Appraisal theories.*** Around the same time as Schachter and Singer were working on their two-factor of emotions, Arnold (1960) developed a similar extension of the Cannon-Bard theory of emotion that included a cognitive component which she termed 'appraisal' (Moors, 2009). Her work was the first of many 'appraisal' theories of emotion and was distinct from the Schachter-Singer model for two reasons. First, her definition of the cognitive component of emotion placed an emphasis on unconscious, automatic cognitive processes in addition to conscious cognition. Second, in her model, the cognitive processing occurred directly after the stimulus and prior to physiological responses (Moors, 2009). In the years that followed, Frijda, Lazarus, Scherer and others have proposed their own versions of Arnold's model, and these theories are usually referred to collectively as appraisal theories of emotion (Moors, 2009). The overall structure of these appraisal theories is as follows: a relevant stimulus in the environment is appraised by the perceiver, and this results in motivation or readiness for action which then may cause physiological responses and/or behavioral responses. Each component contributes to the overall feeling or emotional experience (Moors, 2009).

***Motivational theories – appetitive and defensive systems.*** Lang, Bradley, and Cuthbert (1997) presented a model of emotions as action tendencies. In this model, the authors make use of previous work by Konorski (1967) and Dickinson and Dearing (1979) concerning the two-factor structure of the motivational system (as cited in Cuthbert, 1997). They postulate that the brain has two systems related to motivation- an appetitive and an aversive or defensive system; further, they assert that these two motive systems correspond with the valence dimension of emotion. In this model, unpleasantness or negative valence is associated with the defensive system and pleasantness or positive valence is linked to the appetitive system. The authors propose that the arousal dimension is not specific to either system but instead represents the

extent to which one or both systems is activated (Cuthbert, 1997). Taken together, these two dimensions provide information about which motive system is active (valence) and the intensity of that activation (arousal) (M. M. Bradley, Codispoti, Cuthbert, & Lang, 2001). The authors suggest that even though emotion may be shaped by other factors such as an individual's previous experience or cultural background, the motivational theory provides relatively consistent emotion reactions across the population (M. M. Bradley et al., 2001).

These five theoretical approaches are, by no means, representative of all the theories proposed during the previous century; however, they summarize important historical developments and provide a framework for a continuing discussion of the empirical investigation of emotion, both behaviorally and in the brain.

**Experimental induction of emotion.** An important development in the study of human emotion was the recognition of a need to have effective procedures to induce or manipulate emotion in order to design human experiments. In particular, as researchers became interested in studying the potential interaction of cognitive processes and emotion/mood states, the need for effective emotion induction procedures grew. In the 1970s and 1980s, several types of "mood induction procedures" (MIPs) were developed and utilized (Gerrards-Hesse, Spies, & Hesse, 1994). Goodwin and Williams (1982) reviewed studies involving MIPs designed to induce depression in hopes of gaining a better understanding of clinical depression. In a related vein, Gerrards-Hesse, Spies, and Hesse (Gerrards-Hesse, Spies, & Hesse, 1994) conducted a thorough review of the types of MIPs and their effectiveness at inducing the desired emotion state. The review divided MIPs into the following five groups: (a) MIPs requiring unguided generation of emotion states; (b) MIPs involving guided generation of emotion states; (c) MIPs involving the use of emotional stimuli; (d) MIPs involving engaging participants in emotional situations

related to needs; (e) MIPs involving the generation of physiological states related to specific emotions (Gerrards-Hesse, Spies, & Hesse, 1994). To judge the effectiveness of the different types of MIPs, the reviewers considered evidence from manipulation checks conducted by the experimenters. These are usually conducted directly after the manipulation and aim to measure at least one component of emotion. Generally, the checks measure either the individual's subjective self-report of his/her feeling state, physiological state measures, or observations of behaviors (Gerrards-Hesse, Spies, & Hesse, 1994). Their review found that inducing significant elation was more difficult than inducing significant depression, with 30 percent of MIPs failing to induce significant elation while only 14 percent of MIPs failed to induce depression (Gerrards-Hesse, Spies, & Hesse, 1994). The reviewers also noted that certain subtypes of MIPs had more evidence to support their effectiveness for different mood inductions. Based on the evidence in the paper, the reviewers suggested that if a MIP is going to be used for both elation and depression induction, a subtype of the MIPs involving emotional stimuli (Film/Story MIP) has the best chance for effectiveness (Gerrards-Hesse, Spies, & Hesse, 1994).

**Development of standardized affective visual stimuli.** Early work to categorize MIPs and study their effectiveness provided researchers with a common vocabulary and a greater understanding of the pros and cons of different approaches to mood induction (Gerrards-Hesse, Spies, & Hesse, 1994). However, a significant weakness of early work with MIPs was that even within the same category of MIP, the implementation varied widely (e.g., music or film clips used in one study were different than those used in another study, or the dose of a drug in one study would differ significantly from the dose used in another study)(Gerrards-Hesse, Spies, & Hesse, 1994). The inconsistent MIPs made it difficult to fairly compare results across studies. In response to issues like these, Lang and Bradley developed the International Affective Picture

System (IAPS) (Lang, Bradley, & Cuthbert, 2008) to provide researchers with a large standardized set of visual stimuli to use in the study of emotion and attention. This dataset now contains over 1,000 color photographs across a wide range of categories.

These researchers conceptualized emotion as occurring primarily along two dimensions that correspond to the motivation system engaged (appetitive vs. aversive) and the degree of activation. As such, their approach in developing this dataset was dimensional. In order to measure the affective properties of the images, the developers conducted research in which participants were asked to assess a subset of the images using an affective rating system designed by Lang (1980) called the Self-Assessment Manikin (SAM). Using the SAM, individuals rated images on two major dimensions [affective valence (pleasant to unpleasant) and arousal (excited to calm)] as well as one additional dimension [dominance/control (in control to dominated)] (Lang et al., 2008). Dominance was included as an additional dimension based on Osgood's work on differences in the semantics of emotional evaluations (as cited in Lang et al., 2008). This work found that differences in emotional verbal judgments could be explained by ratings on two primary dimensions, valence and arousal, and one secondary dimension, dominance (Lang et al., 2008). The creators selected images for inclusion in the dataset that covered as much of the range of each dimension as possible (Lang et al., 2008). From participants' SAM ratings, the creators were able to confirm that the images included received consistent ratings, both within and between-subjects and across different experimental settings (Lang et al., 2008). More than a thousand researchers around the world have requested these stimuli from the creators, and the large numbers of studies published that use these stimuli suggest widespread use in behavioral and neuroimaging research. The creators have also developed other standardized affective stimuli including sounds, words, and text (Sander & Scherer, 2009).

**Neuroscientific perspectives.** For over a century, scientists have sought to understand the neural basis of human emotion via indirect methods such as experiments on animals and research with individuals with brain lesions (Adolphs, Tranel, & Damasio, 2003; Cannon, 1927; James, 1884). However, during the last quarter of a century, technological advances have provided scientists with non-invasive methods that allow for more direct study of neural basis of emotion processing in healthy humans (LeDoux, 2000). During this time, hundreds of studies have been conducted using functional neuroimaging techniques to investigate emotion, and making use of the large body of work on emotion in the brain, several researchers have used meta-analytic approaches to test questions about the nature of emotion (L. F. Barrett & Wager, 2006; Hamann, 2012; Lindquist, Wager, Kober, Bliss-Moreau, & Barrett, 2012; Murphy, Nimmo-Smith, & Lawrence, 2003; Phan et al., 2002; Vytal & Hamann, 2010; Wager, Phan, Liberzon, & Taylor, 2003).

Similar to the field of psychology, neuroscience, as a field, does not have a single concise definition of emotion; however, Hamann (2012) proposed a general definition of emotion as a transient affective change that occurs in response to relevant stimuli (external or internal) and involves physiological, neural, and behavioral responses, as well as a conscious experience of the emotion in humans; he also added that these responses often prepare an individual to act (Hamann, 2012).

In 2002, Phan, Wager, Taylor, and Liberzon conducted the first meta-analysis focusing on the functional neuroimaging of emotion using results from 43 PET and 12 fMRI studies of emotion processing in healthy participants. The authors endeavored to define functional neural correlates of emotion by reviewing over a decade of research on emotion. To compare the findings across studies, the authors transformed the significant areas in each study's contrasts of

interest to a standard brain space and then used the transformed activation peaks to compare: (a) regions related to specific emotions; (b) regions related to the method of emotion induction; and (c) regions related to the inclusion or absence of cognitive demand (Phan et al., 2002). No region was consistently active across all studies, which Phan and colleagues suggested was evidence supporting the theory that no single brain region is common to all emotion processing tasks. However, the authors found that the medial prefrontal cortex (mPFC) was activated in many of the studies and that its activation was not related to the specific induction method or emotion type used (Phan et al., 2002). With respect to specific emotion type, the authors discovered that the amygdala was preferentially engaged by fear, the subcallosal cingulate activity was related to sadness, and the basal ganglia appeared to respond significantly to both happiness and disgust. Additionally, the researchers found that the use of a visual induction method (compared to auditory and recall induction methods) was related to greater amygdala activation and the use of imagery or recall induction procedures was associated with anterior cingulate and insula activations. Finally, the researchers also reported that when cognitive demand was added to an emotion task, the anterior cingulate and insula were activated (Phan et al., 2002).

In a more recent meta-analysis, Vytal and Hamann (2010) reviewed PET and fMRI data from emotion studies. The authors aimed to use activation likelihood estimation (ALE) to explore the 'basic' emotion theory. Specifically, the meta-analysis considered the following set of emotions: anger, disgust, fear, sadness, and happiness. Vytal and Hamann (2010) indicated that they chose to use an ALE approach in order to maintain spatial information regarding activation maps in each study and to compare those maps using voxel-wise statistics. This differed from the previously discussed meta-analysis by Phan et al. (2002) which took only the activation peaks from each study, transformed them into standard space, and then compared

across 20 non-overlapping brain regions. Vytal and Hamann identified 83 fMRI and PET studies published between 1993 and 2008 that met their criteria for inclusion in the meta-analyses. The authors conducted two types of analyses to look for evidence of 'basic' emotion in the brain: consistency and discriminability. Consistency analysis revealed areas of activation across studies that were most frequently correlated with each 'basic' emotion. For instance, the ALE method identified clusters in the right superior temporal gyrus (STG) and left anterior cingulate cortex (ACC) as most consistently correlated with happiness and activation in the left medial frontal gyrus, left caudate head, and right inferior frontal gyrus (IFG) as most correlated with sadness. Additionally, the ALE analysis showed significant activation clusters in the left IFG and right parahippocampal gyrus for anger. For fear, the analysis revealed prominent clusters in the bilateral amygdala, right cerebellum and right insula with the largest cluster in the left amygdala; for studies involving disgust, the most consistent activation was found in the bilateral insula with the largest cluster in the right insula and right IFG (Vytal & Hamann, 2010). When looking at areas that were important for discriminability between emotions, the right STG was found to be greater in happiness than sadness and the right middle temporal gyrus (MTG) was greater in sadness than happiness. Similarly, the right STG was found to be greater in comparing happiness and fear while the left amygdala had greater activation in fear than happiness. When comparing happiness and anger, the largest significant cluster for greater activation in happiness than anger was found in the left rostral ACC while the IFG had the largest significant cluster of greater activation in anger than happiness. In a comparison of happiness and disgust, happiness was again shown to have the largest significant cluster in the left rostral ACC while disgust was found to have the largest cluster of activation foci in the right putamen (Vytal & Hamann, 2010). The authors conducted similar pairwise comparisons with all remaining combinations of the five

'basic' emotions of interest and found that each comparison produced a group of brain areas that were capable of reliably discriminating between the pair of emotions; moreover, these patterns of brain activations important for discrimination shared considerable overlap with the areas previously identified as making up the core regions consistently and distinctly activated each 'basic' emotion. The authors argue that this evidence of consistency and discriminability provides good evidence for the theory of 'basic' emotions (Vytal & Hamann, 2010).

**Pattern Classification Approaches to Neuroimaging**

Given the widespread involvement of the brain in emotion processing, information crucial to understanding emotion may exist not only in the brain regions involved in emotion processing but also in the relations among these brain regions. Traditional approaches to neuroimaging data analysis such as statistical parametric mapping (SPM) utilize mass univariate methods (Haufe et al., 2014). These types of analyses examine task-dependent differences in fMRI blood-oxygen-level dependent (BOLD) signal at a voxel-by-voxel level. This approach works well for capturing individual voxels that are significantly different between conditions but is not sensitive to any spatial patterns that may exist among the voxels, as each voxel is treated as an independent test. A multivariate data analytic approach that utilizes pattern classification algorithms may be better suited for the study of emotions as this analysis can capture the richness of spatial information contained in the BOLD signal (Norman, Polyn, Detre, & Haxby, 2006).

Over the last decade, advances in machine learning approaches coupled with improvements in fMRI techniques have led many investigators to consider this approach for analysis of their neuroimaging data. This application of pattern classification algorithms to multi-voxel data sets is known as multi-voxel pattern analysis (MVPA) (Norman et al., 2006).

Norman et al. (2006) provided a review of MVPA approaches, in which they divided the general process into four key steps: (a) "feature selection," (b) "pattern assembly," (c) "classifier training," and (d) "generalization testing" (p. 426). Feature selection is generally used to reduce the amount of data used to build the model by selecting voxels in regions of interest (ROIs) or voxels identified as related to the task condition by univariate analysis. The main goal of feature selection is to reduce noise in order to potentially increase subsequent classifier performance; however, feature selection that results in significant data reduction is not necessary if an accurate classifier can be created from the entire dataset as LaConte et al. (2005) demonstrated with their successful application of MVPA using whole-brain data. The second step, pattern assembly, involves assigning labels to brain patterns at various time points in the task. This process allows brain patterns to be related to the condition of the task that produced the neural activity (Norman et al., 2006). For fMRI data, an important part of this label assignment is accounting for factors such as hemodynamic response delay. The third step is to train a classifier by running a subset of data and corresponding labels through the chosen supervised learning algorithm. This allows the algorithm to learn the optimal function for determining task condition based on the pattern of brain activation. Finally, in the fourth step, the classifier is tested by feeding new data into the classifier and comparing the algorithm's classification with the actual task condition for the data provided (Norman et al., 2006).

MVPA approaches offer multiple benefits including the ability to validate the approach by using performance metrics like prediction accuracy and the suitability of the approach for use in rt-fMRI settings for neurofeedback or brain-computer interfaces (BCI)(LaConte et al., 2005). Haynes and Rees (2006) also enumerate advantages to multivariate neuroimaging analysis. The authors note that there may be information that when judged at an individual voxel level, is too

weak to be regarded as significant, but when the information is aggregated across many voxels using multivariate analyses, the information gains significance. Moreover, they suggest that the useful information may not lie in the knowledge that a single voxel is activated but instead may only carry importance when coupled with knowledge regarding the activation of other voxels (Haynes & Rees, 2006). Additionally, MVPA approaches allow the researcher to make use of detailed spatial and temporal information when needed that is often discarded during traditional mass univariate approaches in order to improve the signal-to-noise ratio such as spatial smoothing and averaging activity across all trials or all blocks of a condition (Haynes & Rees, 2006).

**Support vector machines and their application to neural data.** One supervised learning technique that has gained popularity in neuroimaging analysis is the use of support vector machines (Hollmann et al., 2011; Kamitani & Tong, 2005; LaConte et al., 2005; LaConte, Peltier, & Hu, 2007; Lee, Halder, Kübler, Birbaumer, & Sitaram, 2010; Mourão-Miranda, Bokde, Born, Hampel, & Stetter, 2005; Orrù, Pettersson-Yeo, Marquand, Sartori, & Mechelli, 2012; Sitaram & Veit, 2011). SVM was originally developed by Vapnik (1999) as part of statistical learning theory but has since been applied to a diverse array of problems. SVM approaches are well-suited for use in MVPA because of their distinct ability to handle the pairing of high dimensional data collected over a relatively small sample size and because they appear to be less sensitive to the effects of different preprocessing parameters than other approaches such as canonical variates analysis (CVA) (LaConte et al., 2005).

Generally, when support vector machines are used with fMRI data, neural activity at time *t* is represented as a vector, $\overrightarrow{x_t}$, composed of BOLD signal measurements from all brain voxels (or from the voxels selected during feature selection) and the task condition at time *t* is

represented with a scalar label, $y_t$ (LaConte et al., 2005). SVM uses a non-linear function, $g(\cdot)$: $\vec{z} = g(\vec{x})$, to transform the input vectors, $\vec{x}$, into high dimensional feature space. In the case of linear SVM, $\vec{z}$ is equal to the input vector, $\vec{x}$, as the feature space is the original input space, ($\vec{z} = \vec{x}$). In linear SVM, the algorithm searches for the decision boundary, $\vec{w}$, (referred to as the hyperplane) upon which the two classes are best separated or distinguished, using the decision function (LaConte et al., 2005)

$$D(\vec{z}) = (\vec{w} \cdot \vec{z}) + w_0, \quad (1)$$

This boundary is defined by a subset of the training vectors that were determined to be the most difficult to classify, known as the support vectors (LaConte et al., 2005). Additionally, in order to allow for error in classification of data or to increase the generalizability of the model to new data, a soft-margin can be used (LaConte et al., 2005). In this formulation, the difference between the true class label and the prediction given by the decision function is captured by slack variables, $\xi$. When using slack variables, the decision boundary is defined by $y_t[(\vec{w} \cdot \vec{z}_t) + w_0] \geq 1 - \xi_t$ and is optimized when $\frac{C}{T}\sum_{t=1}^{T}\xi_t + \frac{1}{2}\|\vec{w}\|^2$ is minimized. In this equation, *C* impacts the balance between the complexity of the model and the number of samples that are non-separable (i.e., classified within the margin or on the incorrect side of the margin's boundaries) and the complexity of the model (LaConte et al., 2005).

**Real-Time fMRI**

One application of brain models created using SVM has been using them as a novel and direct avenue for facilitating brain state regulation and studying its effects through real-time fMRI-guided neurofeedback (rt-fMRI). The basic steps of rt-fMRI include collecting fMRI data, performing raw pre-processing on the images, utilizing a machine learning technique to classify data as they are being collected, and finally using the model's classification to provide feedback

to an individual about his/her brain state (i.e., magnitude of activation or direction/rate of change in activation) (deCharms, 2007). Research using this technique has demonstrated that healthy individuals can successfully modulate brain states across multiple domains including motion, visual learning, pain perception, and emotion (Caria, Sitaram, & Veit, 2010; deCharms & Maeda, 2005; LaConte, 2011; Shibata, Watanabe, Sasaki, & Kawato, 2011; Sitaram & Veit, 2011). More recently, researchers have begun testing the use of rt-fMRI with clinical populations as a means of altering abnormal brain responses thought to be related to the disorder or problem (deCharms & Maeda, 2005; Hanlon et al., 2013; Sitaram et al., 2014; Yuan et al., 2014) and have had some success. However, most of these studies involved feedback based on one or more ROIs and have generally provided feedback based on each individual's activity rather than from a normative group model.

The development of a whole-brain, group model of emotion processing would provide a target for emotion regulation and studying the effects of that regulation. LaConte (2007) has adapted SVM software and developed a framework for implementing whole brain feedback in real-time and has demonstrated its promise (LaConte, 2011; LaConte et al., 2007; Papageorgiou, Lisinski, McHenry, White, & LaConte, 2013), and a normative affective model could potentially be translated for use in neurofeedback interventions with individuals with emotion regulation difficulties.

Given the research suggesting emotion dysregulation may be a transdiagnostic marker of psychiatric illness that is related to both symptom severity and treatment response (Fairholme et al., 2013; Werner & Gross, 2010; Wirtz, Radkovsky, Ebert, & Berking, 2014), novel interventions targeting emotion processing could be very fruitful. As initial steps toward the development of such a treatment, this study aimed to (a) develop a whole-brain, group model

that could classify the valence (Positive, Negative, Neutral) of affective brain states using SVMs; (b) optimize the model by testing the impact of varying design choices on the model's performance (c) validate the optimized SVM model using leave-one-out cross validation; (d) adapt the model for use in a real-time fMRI (rt-fMRI) environment and employ an independent dataset to simulate and measure real-time predictions.

## Methods

### Participants and task

**Training dataset**. Fifty-three participants (27 women; mean age 27.2 years; standard deviation (SD) 7.6 years) were recruited from the metropolitan Houston area for this study. Exclusion criteria included left-handedness, contraindications to MR scanning, current use of psychotropic medication, active neurological disease, and any self-reported history of substance dependence or major head trauma. In accordance with the Institutional Review Board of Baylor College of Medicine, after a description of the study's purpose and procedures, written information consent was obtained from all participants. All subjects were compensated monetarily for their participation.

**FMRI task**. During fMRI scanning, participants passively viewed emotional images drawn from the International Affective Picture System (IAPS) (Lang & Cuthbert, 1997). Positive (Pos), negative (Neg), and neutral (Neu) visual stimuli were presented in a block design, pseudo-randomized such that no more than two blocks of the same stimulus type occurred consecutively. In each stimulus block, eight to ten images of the same stimulus type were presented in four-second intervals. The paradigm included 24 stimulus blocks (eight of each stimulus type) separated by jittered, two-to-eight second fixation blocks. Positive and negative stimuli were

selected to include a mixture of high and low arousal images; stimuli were selected such that positive and negative stimuli were matched on mean arousal and mean intensity of valence.

**Real-time simulation dataset**. Thirty-three individuals from the original sample returned to the lab two to four weeks after their initial visit, to complete the fMRI task a second time. The procedures for this visit were identical to the first visit. Due to a data acquisition error, one subject's scanning data was unusable; thus, the final dataset for analysis included 32 individuals (15 women; mean age 27.9 years, SD 7.6 years).

**Data acquisition**

All imaging was performed on 3.0-T Siemens Trio scanners. High-resolution $T_1$-weighted structural scans were collected using the MP-RAGE sequence (Siemens). Whole-brain echo planar images (EPI) hyperangulated 30° from the anterior cingulate-posterior cingulate line were continuously acquired during the passive viewing task. The following scanning parameters were used: repetition time (TR) = 2000 milliseconds; echo time (TE) = 30 milliseconds; flip angle = 90°; 34 axial slices, 4.0 millimeter (mm) slice thickness, 220 x 220 mm field of view, 64 x 64 x 64 matrix with voxel size of 3.4375 x 3.4375 x 4.0 mm.

**Study of model design**

**Preprocessing & first-level analysis of training dataset**. To prepare data for use in SVM training and in-sample testing, imaging data from Visit One were preprocessed using SPM8 (http://www.fil.ion.ucl.ac.uk/spm). Each participant's data underwent realignment, co-registration to his/her structural scan, segmentation, normalization to the Montreal Neurological Institute (MNI)-152 template, smoothing with a 6 mm Gaussian kernel, and correction for six motion parameters. Then, using a first-level general linear model (GLM) that included a constant

and regressors for six motion parameters, subjects' data were detrended and further corrected for motion. A high pass temporal filter of 128 seconds was applied to reduce the effects of low frequency signal drift, and a whole-brain mask restricted analysis to voxels within the brain. The residual images resulting from this GLM analysis were transformed for use in Analysis of Functional NeuroImages (AFNI) (Cox, 1996). Subsequent imaging analysis was performed using AFNI.

**General approach**.

Because our aim was to develop a neural model that could classify brain states associated with passive viewing of Pos, Neg and Neu images, we studied the impact of varying several design options during model development. Specifically, during model construction, we examined the effects of varying the temporal compression of training data and varying the number of individuals used to train the model with. During real-time fMRI simulations, we studied the impact of averaging across real-time input (i.e., brain volumes) and averaging across model classifier output. Because our data included three classes of interest (Pos, Neg, Neu), three separate binary SVM classifiers (Neu vs. Pos, Neu vs. Neg, Pos vs. Neg) were produced for each design version. Having multiple two-class models provided the added benefit of evaluating the differential impact of design choices on multiple options. Although previous work has been conducted concerning how these binary models can be combined to make predictions about three or more classes at a time (Hsu & Lin, 2002; Lorena & de Carvalho, 2004), we limited our investigation to studying the typical two-class SVM model. Future work should consider how design choices impact multiclass predictions as well.

When evaluating the success of different design choices, our primary decision-making metric was prediction accuracy, which we calculated for each binary SVM classifier as

$\left[\frac{number\ of\ correctly\ classified\ examples}{total\ number\ of\ examples}\right] \times 100$. For each of the three binary classifiers, we also calculated the sensitivity of each class as $\left[\frac{number\ of\ true\ positives}{total\ number\ of\ positive\ examples}\right] \times 100$. Additionally, when applicable, we considered other metrics of interest including computational time, model sparseness, interpretability of weight maps, and feasibility of implementation.

Given our desire to design a subject-independent model, or one that could be applied in future experiments to the predict the brain states of previously unseen subjects, our study of affective model parameters focused exclusively on group models (i.e., models that were trained on data from two or more subjects). Our general approach to model design involved randomly selecting two or more individuals from our sample to use for training data and then, selecting one individual from the remaining subject pool to use for testing. To obtain estimates of prediction accuracy for each design option, we repeated this process multiple times and averaged classification performance across all versions.

We chose to use whole-brain data and forego feature selection aimed at limiting the number of voxels included in training. This decision was motivated, in part, by previous success classifying whole-brain data using SVM (LaConte et al., 2005), and also by research showing that in group models, feature selection using an ROI approach, did not affect prediction accuracy (Mourão-Miranda, Reynaud, McGlone, Calvert, & Brammer, 2006). With regards to SVM parameters, we chose to use a linear kernel as LaConte et al. (2005) had previously found that the use of a 2nd or 3rd degree kernel did not appear to confer substantial benefits when using whole-brain training data. Furthermore, this choice had the added benefit that input space was the same as feature space, which allowed for a direct visualization of the SVM weight vector, $\vec{w}$. We used the default *C* parameter of 100 in accordance with LaConte et al.'s (2005) finding that performance accuracy was not significantly impacted by the value of *C* as long as *C* was

sufficiently large (i.e., close to 1 or greater).

**Testing the effects of averaging training data across time**. We evaluated the impact of compressing the training data by averaging across brain volumes collected during each stimulus block. In line with previous work with between-subject models (Mourão-Miranda et al., 2006) and to ensure group models would be computable within a reasonable time frame, we initially used a modest group of 16 subjects as training data for our group models when investigating the impact of temporal compression on model performance. We randomly selected 16 individuals from our participant pool to act as training data; then, from the remaining participant pool, we randomly chose one individual to act as test data. This process was repeated nine times with replacement, resulting in 10 sets of 17 subjects (16 for training and one for testing).

For our uncompressed group models, whole brain volumes (from each individual TR) of the experimental time series of 16 randomly selected subjects were used as training data. To adjust for the hemodynamic delay, the class labels were modified such that the initial two time points of each image block were censored and the two time points immediately following the end of each image block (which were generally collected during a fixation) were included as class examples. To ensure consistent block lengths could be used across all levels of compression, we focused on only the first 16 TRs of each block (as previously defined by the labels shifted to account to hemodynamic delay) and ignored subsequent TRs that were in the stimulus block. We also excluded additional TRs collected during fixation blocks.

Next, we created four compressed training dataset variants by averaging across 2, 4, 8, or 16 adjacent TRs within each stimulus block of the uncompressed training datasets. We altered the class labels to reflect this averaging, such that each stimulus block in the 2 TR compressed

dataset had 8 class labels corresponding to 8 averaged brain volumes, each stimulus block in the 4 TR compressed dataset included 4 class labels corresponding to 4 averaged brain volumes, and so on.

After our training datasets and labels were prepared, we used the 3dsvm plugin (LaConte et al., 2005) for AFNI to train the models. To measure the impact of training data compression on computability of SVM models, we also recorded the computational time required to train each model variant and calculated the mean time required for training a model at each level of compression.

In order to fully understand the impact of training data compression on the model's ability to predict new data, we tested each model on three versions of the same testing dataset. We first tested the models' abilities to predict a previously unseen subject's data on a TR-by-TR basis. Second, we created matching averaged variants of the testing datasets such that within each stimulus block of each test dataset, testing data was averaged across the same number of time points as the training dataset of the model had been. Finally, we evaluated the model variants' abilities to predict testing data that had been averaged across each stimulus block (i.e., all time points within a stimulus block were averaged together to create a single block mean image).

**Testing the effects of varying sample size of training dataset.** To understand how the size of the group used to train the model impacts model performance, we trained models using six different group sizes. We started with the smallest group possible (two subjects), and for each new variant, we doubled the group size (i.e., four subjects, eight subjects, 16 subjects, and 32 subjects). For our sixth group size, we used 52 subjects (i.e., one less than our total sample size). To obtain estimates of prediction accuracies, we created multiple versions of each model size.

Specifically, since our 52-subject model required the use of all 53 subjects in the sample, either for training or testing, only 53 unique versions of this model existed. Thus, we also created 53 versions of each smaller group size. Similar to the methods used to create our previous group models, we randomly selected two, four, eight, 16, or 32 subjects from our sample to use as training data and then randomly selected one additional subject from the remaining subject pool to use as testing data. This was done 53 times with replacement.

Based on the results obtained in the previous step, we decided to use training data that had been averaged across all time points within a stimulus block. Thus, each subject contributed 24 images (8 of each class) to their group models. To prepare each training dataset, the selected subjects' data were concatenated together and the corresponding class labels were also concatenated. The models were trained using the 3dsvm plugin for AFNI. In line with our previous finding that prediction accuracies were highest when models were predicting on block mean data, the test subjects' data were averaged across all time points within each block to create block mean images for testing. The trained models were each used to predict the brain states of a test subject, and model performance was averaged across all 53 versions of each group size to obtain an estimate of prediction accuracy.

Additionally, we examined two secondary metrics to better understand the impact of training data sample size on model health. First, we calculated the percentage of training data identified as support vectors for each model and averaged across the 53 results for each group size. Second, we examined the $\vec{w}$ maps for each group size, as well as the maps' distributions of voxel weights within each map. To facilitate this comparison, within each group size, $\vec{w}$ maps from each of the 53 model versions were pooled and a t-test was performed to identify the average estimated patterns of weighted voxels. The resulting maps were z-scored and

thresholded to display on the top 10% of weighted voxels identified for each group size. Histograms of the distribution of weighted voxels for each group size were plotted with the 10% threshold marked with dashed lines.

**Validating group model.** The 52-subject model training and testing described previously also served as a way to test the validity of the overall 53-subject group model as it adhered to the leave-one-out cross validation method. For each iteration of the model, a single subject's dataset was left out (24 images), the remaining 52 individuals' datasets were used to train a 52-subject group model, and then the model was tested on the left-out subject's data. This was repeated for each subject, and by averaging across the corresponding 53 "validation" prediction accuracies, estimates of the 53-subject group model's performance on each of the three, two-class SVM comparisons were obtained.

**Translation of group model for use in rt-fMRI simulations.** Because of our interest in the translation of the model for future use as a neurofeedback tool, we investigated our model's performance in a simulated rt-fMRI environment using previously acquired neuroimaging data. We conducted these real-time simulations using 3dsvm, (LaConte, 2011; LaConte et al., 2005; 2007) via the command line. At the start of each simulation, un-preprocessed raw imaging data were imported, functional images were deobliqued, and a mask was created using the first 3 TRs of data. Then, to prepare the group model for use with new subjects, we added a step to transform the group model into the subject's native space. Specifically, we aligned the first four volumes of a subject's EPI sequence to an EPI template in MNI space, and the transformation matrix for this alignment was saved. Then, the inverse of this transformation matrix was calculated and applied to the group model in order to transform the group model from MNI space into each subject's native space. We chose to transform the group model from MNI space to a

subject's native space instead of transforming the subject's imaging data to standard MNI space because it enabled us to perform a single transformation at the start of the experiment rather than needing a transformation at every TR in order to change a subject's data from native space to MNI space (LaConte, 2011).

Following this transformation, the simulation proceeded as usual, applying minimal preprocessing to the incoming data to correct for motion and classifier drift (LaConte et al., 2007) and then making brain state predictions on a TR-by-TR basis using our group model. This process was repeated for each of the remaining 31 subjects, and estimates of the model performance were obtained by averaging the results across the group.

**Testing effects of averaging real-time fMRI inputs (brain volumes).** To explore potential methods of improving reliability of the rt-fMRI feedback signal, we tested several simulation variations. After testing the model initially on uncompressed data for which a separate prediction was made for every TR, we studied the impact of averaging the inputs (i.e., brain volumes) prior to submitting them to the model for prediction. We tested the effects of averaging across two, four, eight, and 16 time points using a moving window method. In this approach, the SVM model made a prediction every TR; however, the brain volumes the model was predicting on had been averaged across TRs in the window. For example, in the case of a four TR moving window, at the first time point, the model made a prediction based on a single brain volume; at the second time point, the model's prediction was based the average of the first two brain volumes; starting with the fourth time point, the model's prediction was based on the average of the current TR and the three TRs prior to the time point. We conducted two types of simulations with this moving window technique: (a) continuously across the experimental time course and (b) continuously only within each block (i.e., starting a new window at the beginning

of each new block). Model performance of each version was estimated by averaging across results from the 32 participants.

**Testing effects of averaging outputs (model classifier outputs).** Additionally, we tested a second method aimed at reducing noise and increasing the reliability of the model's predictions: averaging the classifier output across multiple time points. Using the same moving window method of averaging we used when averaging inputs, we tested the impact of averaging the model classifier output across two, four, eight, and 16 TRs. We tested this moving block window continuously across the time course as well as solely within each block. Estimates of overall model performance were obtained by averaging across the results of the individuals in the sample.

**Time course of model performance across block.** Additionally, to better understand how classifier averaging impacts predictions across the time course of a block, we calculated the average prediction accuracy at each TR across the times series of the block.

## Results

### Temporal Compression of Training Data

**Effect on computational time to train models.** On average, training a group model on all three stimuli classes (Pos, Neg, and Neu) using uncompressed data from 16 subjects required close to four days to compute (mean = 93.20 hours; SEM = 5.83 hours). When the neuroimaging data from each subject was compressed by averaging across sets of two TRs, on average, training a group model using data from 16 subjects took less than one day to calculate (mean = 17.08 hours; SEM = 0.80 hours). As Figure 1A depicts, further compression of the training data continued to result in decreased model training time, such that models trained on data

compressed across sets of four, eight, and 16 TRs had average training times of approximately four hours, one hour, and 17 minutes respectively.

**Effect on model performance when testing on uncompressed data.** Compression of training data did not adversely impact average model performance when tested on uncompressed data. In fact, greater degrees of within-block compression of training resulted in improvements in prediction accuracy when tested on uncompressed data (left panel, Figure 1B). Paired t-test comparisons of accuracies obtained from each temporal compression version of the model showed that the 16 TR compressed model performed significantly better than the 1 TR model ($p < .05$) for both the Neu vs. Pos classifier and Pos vs. Neg classifier. For the Neu vs. Neg classifier, the 16 TR compressed model performed significantly better than any other model version (all paired comparisons with 16 TR model: $p < .001$). The Neu vs. Neg SVM classifier was significantly more accurate when predicting uncompressed training data than either the Neu vs. Pos SVM or the Pos vs. Neg SVM. Looking at the class sensitivities (left panel, Figure 1C), difficulty detecting the Pos class (i.e., under 55% sensitivity) seemed to be driving the reduced prediction accuracies of both the Neu vs. Pos SVM and the Pos vs. Neg SVM.

**Effect on model performance when testing on matching averaged data.** Compression of training data resulted in significantly greater prediction accuracies for test data that had been similarly compressed (i.e., for a model trained on data averaged across 4 TRs, testing data was also averaged across 4 TRs), as evidenced by paired t-test comparisons of prediction accuracies for each model version (middle panel, Figure 1B). Notably, paired comparisons of all other compression sizes to the 16 TR model were significant ($p < .001$) for all three binary classifiers. The Neu vs. Neg SVM classifier again had higher prediction accuracies than either the Neu vs. Pos SVM or the Pos vs. Neg SVM. Examining the class sensitivities (middle panel, Figure 1C),

the Pos class again was the most difficult to detect for both the Neu vs. Pos SVM and the Pos vs. Neg SVM. However, compression of training data did improve both classifiers' sensitivity to the Pos class when testing on matching averaged data.

**Effect on model performance when testing on block mean data.** Compression of training data did not adversely impact prediction accuracies on block mean testing data, as evidenced by paired t-test comparisons that showed either significant improvements due to compression or no differences due to compression (such as in the Pos vs. Neg binary classifier), but not significant decrease in performance. For the Neu vs. Neg SVM classifier, compression up to 16 TRs led to gains in model performance, as supported by a paired t-tests indicating a significant difference between accuracies in the 1 TR model version and in the 16 TR model version ($p < .05$) (right panel, Figure 1B). Once again, the Neu vs. Neg SVM classifier showed significantly better prediction accuracy than either the Neu vs. Pos or the Pos vs. Neg SVM classifiers, and this, again, looked to be due to problems with the detection of the Pos class for both the Neu vs. Pos SVM and the Pos vs. Neg SVM (right panel, Figure 1C).

**Sample Size of Model Training Group**

**Effect on percentage of data identified as support vectors.** As the size of the training group increased, the percentage of data identified as support vectors decreased (Figure 2A). This suggests the model became sparser as the sample size of the training data grew because the percentage of the training data that either defined the margin boundary or was non-separable was significantly less. Notably, when models have a very high percentage of data identified as support vectors, it may indicate over-fitting of a model (Cortes & Vapnik, 1995). Regardless of the training group size, in the Neu vs. Neg SVM classifier, the percentage of support vectors was

significantly lower than either the Neu vs. Pos SVM classifier or the Pos vs. Neg SVM classifier, indicating it may be easier to discriminate between Neu and Neg.

**Effect on model performance when testing on block-averaged data.** Increasing the number of individuals whose data were included in training improved the accuracy of the models. However, once the training dataset was sufficiently large (approximately 8 to 16 subjects), the improvements in model performance were negligible. For example, examining the Neu vs. Pos classifier, paired t-tests showed that the 8-subjct, 16-subject, 32-subject, and 52-subject models were all significantly better than the 2-subject or 4-subject models (all $p < .05$) but were not significantly different from each other. Consistent with the results in Figure 1, the Neu vs. Neg SVM classifier performed significantly better than either the Neu vs. Pos classifier or the Pos vs. Neg classifier, and the lag in performance of the Neu vs. Pos and Pos vs. Neg classifiers appeared to be related to the model's difficulty in detecting the Pos class.

**Effect on interpretability of SVM Neu vs. Neg weight map.** Overall, several brain regions appeared consistently across group sizes including portions of the thalamus, hippocampus, and parahippocampal gyrus as well as areas in the occipital lobe and medial prefrontal cortex (mPFC). As the sample size of our training group increased from two subjects up to 16 subjects, the average Neu vs. Neg classifier map showed the addition of brain regions previously implicated in emotion processing (e.g., insula), suggesting that with more training subjects, these areas appear more reliably. However, as sample size continued to increase to 32 subjects and 53 subjects, the maps became increasingly sparse with many regions being represented by a smaller number of voxels. While many of the small clusters displayed in the larger group maps are still in areas implicated in emotion processing, the decreasing size of the

clusters as well as the increasing number of small clusters of activity across the brain limit the clear interpretability of the map.

**Effect on distribution of voxel weights.** In the histograms of models trained on smaller group sizes, there appeared to be significantly more heavily weighted voxels with positive values than with negative values. This trend seemed to normalize as the sample size of the group increased, with the 52-subject group size appearing to have approximately the same distribution of positive voxel weights as negative voxel weights.

**Group model validation.** Using leave-one-out cross-validation, we estimated the prediction accuracy of the 53-subject group model. As Table 1 shows, all three binary classifiers had a mean accuracy above 70%, with the Neu vs. Neg SVM classifier showing the best performance with a mean accuracy of 87.62%.

**Translation of group model for use in rt-fMRI simulations**

**Effects of averaging real-time fMRI inputs (brain volumes).** When averaging across brain volumes continuously, prediction accuracy was the best when averaging four TRs (Figure 4A, left). However, when averaging across brain volumes within block only, as more brain volumes were averaged together, prediction accuracy improved, with the best accuracy when averaging brain volumes across 16 TRs within block (Figure 4A, right). In both conditions, the Neu vs. Neg SVM classifier had the best prediction accuracy, the Neu vs. Pos SVM classifier performed the second best, and the Pos vs. Neg SVM classifier performed the worst.

**Effects of averaging outputs (model classifier outputs).** When averaging across the classifier outputs continuously, prediction accuracy increased as the number of TRs averaged increased from two TRs to eight TRs and leveled off for 16 TRs (Figure 4B, left). When averaged within block only, improvements were seen in prediction accuracy for averaging up to

16 TRs (Figure 4B, right). Similar to the results based on averaging brain volumes, the Neu vs. Neg SVM classifier performed the best and the Pos vs. Neg SVM classifier performed the worst.

**Comparison of the averaging methods.** Overall, averaging using a moving window within the block only produced the best model performance, and averaging the classifier output resulted in better prediction accuracy than averaging brain volumes.

**TR-by-TR accuracy across average block time series.** As Figure 4C depicts, when predicting at every TR without averaging, the average prediction accuracy dropped significantly as the time course of the block continued. Averaging the classifier output across a moving window of 16 TRs resulted in relatively stable prediction accuracies with only a slight loss of accuracy during the last four TRs of the block.

## Discussion

Here, we created a group model of brain states related to the passive viewing of emotional stimuli, and we showed that offline model performance was optimized by (a) averaging training data across time points within each block and (b) including larger numbers of individuals in the training dataset. Furthermore, we demonstrated that the resulting group model could reliably classify brain states related to Pos, Neg, and Neu stimuli. Finally, we demonstrated one method of translating this group model for use in rt-fMRI, and through simulations using previously collected neuroimaging data, we tested several methods of making predictions on a TR-by-TR basis.

Although averaging across TRs in a block is one method for reducing computational cost and potentially noise (Mourão-Miranda et al., 2006), it was unclear if this temporal compression would limit a group model's ability to make predictions on uncompressed training data, such as in rt-fMRI. Our findings indicated that temporal compression not only preserved prediction

accuracy, but also, in many instances improved accuracy of the model when predicting on uncompressed data. This suggests that patterns of activation essential to distinguishing between affective states may have been relatively stable across the block time course. However, temporal compression may have varying effects depending on the time course of the neural response being predicted. Although, in our results, averaging did not lead to significant decreases in performance accuracy, the Pos class appeared to show very little benefit from averaging training data. This may be related to a more variable neural response across the time course of positive images or may be the result of Pos being a weaker neural response that is difficult to distinguish at the level of individual TRs.

With regards to the size of group needed to train a reliable model, we demonstrated that inclusion of data from additional subjects in the training dataset improved the model accuracy. Surprisingly, group models built on only two subjects' data were able to predict brain states of new individuals significantly better than chance, with two-subject models showing an average accuracy of 77.83% when distinguishing between Neu and Neg classes. This finding points to the robust, consistent patterns of activation elicited by emotional stimuli. While initially increasing the sample size of the data led to increasing accuracy, this effect appeared to plateau around 16 subjects, suggesting a relatively modest sized group is sufficient in creating group models of brain states. Because we only have estimates of the model's prediction accuracy, it is unclear if the addition of subjects beyond 16 resulted in very small increases in accuracy that may only be reliably detected if tested on a larger sample or if it results in no improvement at all, as our overlapping SEM bars suggested. Additionally, it is unknown if much larger increases in sample size, say the inclusion of 100 or more subjects, would lead to notable increases in accuracy. To our knowledge, there has been only one study with a sample size in that range that

used a comparable emotion elicitation paradigm used to create a multivariate group model, which trained a group model on 121 subjects (Chang, Gianaros, Manuck, Krishnan, & Wager, 2015). Their study focuses on predicting the intensity of Neg valence rather than distinguishing between Neg and Neu classes, so they reported the average correlation between participant-rated intensity and the predicted intensity. Their results on the impact of sample size on prediction accuracy are remarkably similar to our findings; training with only 16 subjects, they reported an average correlation of approximately 0.8 and when using 52 subjects, the reported correlation is 0.9. For the model trained with 121 subjects, the average correlation is 0.92 with a standard deviation of 0.01 (Chang et al., 2015). Taken in conjunction with our results, it appears that for predictions related to comparing Neu and Neg images, at least, prediction accuracy of group models are high even with sample sizes of less than 20 subjects, and the addition of more subjects may provide relatively small increases in prediction accuracy. However, it appears that once the sample size is sufficiently large (more than 50 subjects) or the prediction accuracy is high (near 90%), even large increases in sample size provide negligible improvements in accuracy.

One consistent result that we observed across model versions was that the prediction accuracy of Neu vs. Neg classifiers appeared to be significantly better than the accuracy of either the Neu vs. Pos classifier or the Pos vs. Neg classifier. Interestingly, this discrepancy seemed to be, in large part, the result of difficulty detecting Pos stimuli, as was shown when graphing the sensitivity of each class separately. Previous research has found that positive emotion is harder to induce experimentally than negative emotion, with approximately 30% of Pos manipulations failing to result in significantly mood differences and only about 14% of Neg manipulations being unsuccessful at inducing significant mood differences (Gerrards-Hesse, Spies, & Hesse,

1994). Thus, one reason for poorer prediction accuracy of Pos states may be that the Pos images failed to induce a Pos mood in some. Additionally, some research indicates that particular regions of the brain are differentially modulated by the interaction of arousal and valence, with the occipital cortex showing early activation to Neg stimuli regardless of arousal but showing early activation to Pos stimuli only when the stimuli are also rated high arousal (Nielen et al., 2009). It is possible that interactions such as these may have contributed to the difficulty classifying positive stimuli as we included a mix of high and low arousal images. Alternately, researchers have shown a relation between positive mood induction and the widening of attentional focus, generally to include irrelevant information in the environment (Biss & Hasher, 2011); in the context of our experiment, if an individual's attention was broadened during positive image blocks, it could have led to encoding of extraneous information from the environment, which could have increased noise in the Pos class.

Examination of the $\vec{w}$ maps produced by the Neu vs. Neg SVM classifier found consistency across multiple regions regardless of sample size, suggesting the robustness of the pattern of brain activity elicited. As sample size increased from two subjects to 16 subjects, weight maps identified new areas that have previously been implicated in emotion processing, such as the insula, suggesting that increased sample size does help with the reliable detection of significant brain regions. However as the group size continued to increase, the sparseness of the weight maps increases, with important regions being represented by fewer voxels. This renders the 32 and 53 subject maps less easily visually interpretable. However, the histograms of the weight distributions of voxels suggested that these larger sample sizes have more equal distribution of positive and negative weights, which may indicate that each class is represented more stably.

With regards to our real-time fMRI simulation, we demonstrated that a group model could be successfully translated for use with new subjects by calculating the transformation from each subject's native space to standard MNI space and then applying the inverse of this transformation to the group model. Then, we showed that the group model could provide reliable feedback on a TR-by-TR basis by employing a method of within-block moving window averaging across 16 TRs of classifier output. It appears that the moving window averaging helped reduce model errors due to signal drop off that occurred over the time course of the block. This result reinforces the importance of attending to the temporal dynamics of the predicted brain states. This type of averaging may be unnecessary if the length of blocks were briefer or if the state being modeled was one with a more prolonged response.

Using the moving window averaging during our rt-fMRI simulations, we were able to achieve average prediction accuracies that were significantly above chance and better than the offline prediction accuracy of our 16-subject model predicting on unaveraged data (Figure 1B). However, only the Neu vs. Neg classifier had average prediction accuracy above 70%. As such, providing TR-by-TR feedback for the Neu vs. Pos classifier or the Pos vs. Neg classifier with this version of the model may be suboptimal. Future work should investigate the optimal trade-off between frequency of feedback and optimization of prediction accuracy. Additionally, in the future, exploration of different methods of transforming group models for use in real-time fMRI as well as the impact of preprocessing parameters on these transformations could provide valuable information that would enable further optimization of prediction accuracy.

## Conclusion

In summary, our study found that it is possible to create a whole-brain group model of affective brain states that can reliably classify brain states of new individuals and can also be

transformed for use in a real-time setting. The model was most accurate when distinguishing between Neg and Neu classes and appeared to have the most difficulty detecting Pos brain states; however, all three binary classifiers performed significantly above chance in both offline analyses and real-time simulation. These results show promise for the potential use of a whole-brain group model in a real-time setting to provide normative feedback regarding healthy emotional states, particularly with regards to Neu vs. Neg brain states. Further work is needed to test the generalization of this model to brain states elicited by other methods of mood induction, as well as to explore the impact of this type of feedback during emotion regulation when used with healthy populations prior to testing the model's utility as an intervention for emotion dysregulation in individuals with psychopathology.

## References


Adolphs, R., Tranel, D., & Damasio, A. R. (2003). Dissociable neural systems for recognizing emotions., *52*(1), 61–69. http://doi.org/10.1016/S0278-2626(03)00009-5

Arnold, M. B. (1960). Emotion and personality (Vol. 1). Columbia University Press.

Barrett, L. F., & Wager, T. D. (2006). The Structure of Emotion. Evidence From Neuroimaging Studies. *Current Directions in Psychological Science*, *15*(2), 79–83. http://doi.org/10.1111/j.0963-7214.2006.00411.x

Beauregard, M., Paquette, V., & Lévesque, J. (2006). Dysfunction in the neural circuitry of emotional self-regulation in major depressive disorder. *Neuroreport*, *17*(8), 843–846. http://doi.org/10.1097/01.wnr.0000220132.32091.9f

Biss, R. K., & Hasher, L. (2011). Delighted and distracted: positive affect increases priming for irrelevant information. *Emotion*, *11*(6), 1474–1478. http://doi.org/10.1037/a0023855

Bradley, B., DeFife, J. A., Guarnaccia, C., Phifer, J., Fani, N., Ressler, K. J., & Westen, D. (2011). Emotion dysregulation and negative affect: association with psychiatric symptoms. *The Journal of Clinical Psychiatry*, *72*(5), 685–691. http://doi.org/10.4088/JCP.10m06409blu

Bradley, M. M., Codispoti, M., Cuthbert, B. N., & Lang, P. J. (2001). Emotion and motivation I: defensive and appetitive reactions in picture processing., *1*(3), 276–298. http://doi.org/10.1037//1528-3542.1.3.276

Britton, J. C., Taylor, S. F., Sudheimer, K. D., & Liberzon, I. (2006). Facial expressions and complex IAPS pictures: common and differential networks. *NeuroImage*, *31*(2), 906–919. http://doi.org/10.1016/j.neuroimage.2005.12.050

Cannon, W. B. (1927). The James-Lange Theory of Emotions: A Critical Examination and an

Alternative Theory. *The American Journal of Psychology*, *39*(1/4), 106–124. http://doi.org/10.2307/1415404?ref=no-x-route:2fb30a725f7e82f18b7d5818065027e3

Caria, A., Sitaram, R., & Veit, R. (2010). Volitional control of anterior insula activity modulates the response to aversive stimuli. A real-time functional magnetic resonance imaging study. *Biological Psychiatry*, *68*(5), 425–432. http://doi.org/10.1016/j.biopsych.2010.04.020

Chang, L. J., Gianaros, P. J., Manuck, S. B., Krishnan, A., & Wager, T. D. (2015). A Sensitive and Specific Neural Signature for Picture-Induced Negative Affect. *PLOS Biology*, *13*(6), e1002180–e1002180. http://doi.org/10.1371/journal.pbio.1002180

Cisler, J. M., James, G. A., Tripathi, S., Mletzko, T., Heim, C., Hu, X. P., et al. (2013). Differential functional connectivity within an emotion regulation neural network among individuals resilient and susceptible to the depressogenic effects of early life stress. *Psychological Medicine*, *43*(3), 507–518. http://doi.org/10.1017/S0033291712001390

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. http://doi.org/10.1007/BF00994018

Cox, R. W. (1996). AFNI: Software for Analysis and Visualization of Functional Magnetic Resonance Neuroimages. *Computers and Biomedical Research*, *29*(3), 162–173. http://doi.org/10.1006/cbmr.1996.0014

Cuthbert, B. N. (1997). Motivated Attention. In *Attention and Orienting* (pp. 95–135). Psychology Press.

Dalgleish, T. (2009). James-Lange Theory. In K. R. Scherer & D. Sander (Eds.), *The Oxford companion to emotion and the affective sciences sciences* (p. 229). Oxford ; New York : Oxford University Press.

Dalgleish, T. (2010). Cannon-Bard Theory. In *Handbook of Emotions* (p. 83). Guilford Press.

Dalgleish, T., Dunn, B. D., & Mobbs, D. (2009). Affective Neuroscience: Past, Present, and Future. *Emotion Review*, *1*(4), 355–368. http://doi.org/10.1177/1754073909338307

deCharms, R. C. (2007). Reading and controlling human brain activation using real-time functional magnetic resonance imaging. *Trends in Cognitive Sciences*, *11*(11), 473–481. http://doi.org/10.1016/j.tics.2007.08.014

deCharms, R. C., & Maeda, F. (2005). Control over brain activation and pain learned by using real-time functional MRI. Presented at the Proceedings of the …. http://doi.org/10.1073/pnas.0505210102

Dixon-Gordon, K. L., Tull, M. T., & Gratz, K. L. (2014). Self-injurious behaviors in posttraumatic stress disorder: an examination of potential moderators. *Journal of Affective Disorders*, *166*, 359–367. http://doi.org/10.1016/j.jad.2014.05.033

Etkin, A., & Schatzberg, A. F. (2011). Common abnormalities and disorder-specific compensation during implicit regulation of emotional processing in generalized anxiety and major depressive disorders. *The American Journal of Psychiatry*, *168*(9), 968–978. http://doi.org/10.1176/appi.ajp.2011.10091290

Fairholme, C. P., Nosen, E. L., Nillni, Y. I., Schumacher, J. A., Tull, M. T., & Coffey, S. F. (2013). Sleep disturbance and emotion dysregulation as transdiagnostic processes in a comorbid sample. *Behaviour Research and Therapy*, *51*(9), 540–546. http://doi.org/10.1016/j.brat.2013.05.014

Frijda, N. H. (2010). The Psychologists' Point of View. In *Handbook of Emotions* (pp. 68–87). Guilford Press.

Gerrards-Hesse, A., Spies, K., & Hesse, F. W. (1994). Experimental inductions of emotional states and their effectiveness: A review. *British Journal of Psychology*, *85*(1), 55–78.

http://doi.org/10.1111/j.2044-8295.1994.tb02508.x

Goodwin, A. M., & Williams, J. M. (1982). Mood-induction research--its implications for clinical depression. *Behaviour Research and Therapy*, *20*(4), 373–382. http://doi.org/10.1016/0005-7967(82)90097-3

Hamann, S. (2012). Mapping discrete and dimensional emotions onto the brain: controversies and consensus. *Trends in Cognitive Sciences*, *16*(9), 458–466. http://doi.org/10.1016/j.tics.2012.07.006

Hanlon, C. A., Hartwell, K. J., Canterberry, M., Li, X., Owens, M., Lematty, T., et al. (2013). Reduction of cue-induced craving through realtime neurofeedback in nicotine users: The role of region of interest selection and multiple visits. *Psychiatry Research: Neuroimaging*, *213*(1), 79–81. http://doi.org/10.1016/j.pscychresns.2013.03.003

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., & Bießmann, F. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage*, *87*, 96–110. http://doi.org/10.1016/j.neuroimage.2013.10.067

Haynes, J.-D., & Rees, G. (2006). Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, *7*(7), 523–534. http://doi.org/10.1038/nrn1931

Hollmann, M., Rieger, J. W., Baecke, S., Lützkendorf, R., Müller, C., Adolf, D., & Bernarding, J. (2011). Predicting decisions in human social interactions using real-time fMRI and pattern classification. *PLoS ONE*, *6*(10), e25304–e25304. http://doi.org/10.1371/journal.pone.0025304

Hsu, C.-W., & Lin, C.-J. (2002). A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, *13*(2), 415–425. http://doi.org/10.1109/72.991427

James, W. (1884). What is an emotion? *Mind*, *9*(34), 188–205.

Joormann, J., & Goodman, S. H. (2014). Transdiagnostic processes in psychopathology: in memory of Susan Nolen-Hoeksema. *Journal of Abnormal Psychology*, *123*(1), 49–50. http://doi.org/10.1037/a0035525

Kamitani, Y., & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience*, *8*(5), 679–685. http://doi.org/10.1038/nn1444

LaConte, S. M. (2011). Decoding fMRI brain states in real-time. *NeuroImage*, *56*(2), 440–454. http://doi.org/10.1016/j.neuroimage.2010.06.052

LaConte, S. M., Peltier, S. J., & Hu, X. P. (2007). Real-time fMRI using brain-state classification. *Human Brain Mapping*, *28*(10), 1033–1044. http://doi.org/10.1002/hbm.20326

LaConte, S. M., Strother, S. C., Cherkassky, V. L., Anderson, J., & Hu, X. (2005). Support vector machines for temporal classification of block design fMRI data. *NeuroImage*, *26*(2), 317–329. http://doi.org/10.1016/j.neuroimage.2005.01.048

Lang, P. J. (2010). Emotion and motivation: Toward consensus definitions and a common research purpose. *Emotion Review*. http://doi.org/10.1177/1754073910361984

Lang, P. J., & Cuthbert, B. N. (1997). *International Affective Picture System (IAPS)*.

Lang, P. J., Kozak, M. J., Miller, G. A., Levin, D. N., & McLean, A. (1980). Emotional imagery: conceptual structure and pattern of somato-visceral response. *Psychophysiology*, *17*(2), 179–192.

LeDoux, J. E. (2000). Emotion circuits in the brain. *Annual Review of Neuroscience*, *23*(1), 155–184. http://doi.org/10.1146/annurev.neuro.23.1.155

Lee, S., Halder, S., Kübler, A., Birbaumer, N., & Sitaram, R. (2010). Effective functional

mapping of fMRI data with support-vector machines. *Human Brain Mapping*, *31*(10), 1502–1511. http://doi.org/10.1002/hbm.20955

Linden, D. E. J. (2014). Neurofeedback and networks of depression. *Dialogues in Clinical Neuroscience*, *16*(1), 103–112.

Lindquist, K. A. K., Wager, T. D. T., Kober, H. H., Bliss-Moreau, E., & Barrett, L. F. (2012). The brain basis of emotion: a meta-analytic review. *Behavioral and Brain Sciences*, *35*(3), 121–143. http://doi.org/10.1017/S0140525X11000446

Lorena, A. C., & de Carvalho, A. C. P. L. F. (2004). Comparing Techniques for Multiclass Classification Using Binary SVM Predictors. In A. C. R. Paiva, R. Prada, & R. W. Picard (Eds.), *Affective Computing and Intelligent Interaction* (Vol. 2972, pp. 272–281). Berlin, Heidelberg: Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-540-24694-7_28

Moors, A. (2009). Theories of emotion causation: A review. *Cognition & Emotion*, *23*(4), 625–662. http://doi.org/10.1080/02699930802645739

Mourão-Miranda, J., Bokde, A. L. W., Born, C., Hampel, H., & Stetter, M. (2005). Classifying brain states and determining the discriminating activation patterns: Support Vector Machine on functional MRI data. *NeuroImage*, *28*(4), 980–995. http://doi.org/10.1016/j.neuroimage.2005.06.070

Mourão-Miranda, J., Reynaud, E., McGlone, F., Calvert, G., & Brammer, M. J. (2006). The impact of temporal compression and space selection on SVM analysis of single-subject and multi-subject fMRI data. *NeuroImage*, *33*(4), 1055–1065. http://doi.org/10.1016/j.neuroimage.2006.08.016

Murphy, F. C., Nimmo-Smith, I., & Lawrence, A. D. (2003). Functional neuroanatomy of emotions: A meta-analysis. *Cognitive, Affective, & Behavioral Neuroscience*, *3*(3), 207–233.

Nielen, M. M. A., Heslenfeld, D. J., Heinen, K., Van Strien, J. W., Witter, M. P., Jonker, C., & Veltman, D. J. (2009). Distinct brain systems underlie the processing of valence and arousal of affective pictures., *71*(3), 387–396. http://doi.org/10.1016/j.bandc.2009.05.007

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, *10*(9), 424–430. http://doi.org/10.1016/j.tics.2006.07.005

Orrù, G., Pettersson-Yeo, W., Marquand, A. F., Sartori, G., & Mechelli, A. (2012). Using Support Vector Machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review. *Neuroscience & Biobehavioral Reviews*, *36*(4), 1140–1152. http://doi.org/10.1016/j.neubiorev.2012.01.004

Papageorgiou, T. D., Lisinski, J. M., McHenry, M. A., White, J. P., & LaConte, S. M. (2013). Brain-computer interfaces increase whole-brain signal to noise. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(33), 13630–13635. http://doi.org/10.1073/pnas.1210738110

Phan, K. L., Wager, T., Taylor, S. F., & Liberzon, I. (2002). Functional neuroanatomy of emotion: a meta-analysis of emotion activation studies in PET and fMRI. *NeuroImage*.

Ruiz, S., Buyukturkoglu, K., Rana, M., Birbaumer, N., & Sitaram, R. (2014). Real-time fMRI brain computer interfaces: self-regulation of single brain regions to networks. *Biological Psychology*, *95*, 4–20. http://doi.org/10.1016/j.biopsycho.2013.04.010

Sander, D., & Scherer, K. R. (Eds.). (2009). The Oxford companion to emotion and the affective sciences sciences. New York: Oxford University Press.

Schachter, S., & Singer, J. E. (1962). Cognitive, social, and physiological determinants of emotional state. *Psychological Review*, *69*, 379–399. http://doi.org/10.1037/h0046234

Schlochtermeier, L. H., Kuchinke, L., Pehrs, C., Urton, K., Kappelhoff, H., & Jacobs, A. M. (2013). Emotional picture and word processing: an FMRI study on effects of stimulus complexity. *PLoS ONE*, *8*(2), e55619. http://doi.org/10.1371/journal.pone.0055619

Shibata, K., Watanabe, T., Sasaki, Y., & Kawato, M. (2011). Perceptual learning incepted by decoded fMRI neurofeedback without stimulus presentation. *Science (New York, N.Y.)*, *334*(6061), 1413–1415. http://doi.org/10.1126/science.1212003

Sitaram, R., & Veit, R. (2011). Real-time support vector classification and feedback of multiple emotional brain states. *NeuroImage*, *56*(2), 753–765. http://doi.org/10.1016/j.neuroimage.2010.08.007

Sitaram, R., Caria, A., Veit, R., Gaber, T., Ruiz, S., & Birbaumer, N. (2014). Volitional control of the anterior insula in criminal psychopaths using real-time fMRI neurofeedback: a pilot study. *Frontiers in Behavioral Neuroscience*, *8*, 344. http://doi.org/10.3389/fnbeh.2014.00344

Stoeckel, L. E., Garrison, K. A., Ghosh, S., Wighton, P., Hanlon, C. A., Gilman, J. M., et al. (2014). Optimizing real time fMRI neurofeedback for therapeutic discovery and development. *NeuroImage: Clinical*, *5*, 245–255. http://doi.org/10.1016/j.nicl.2014.07.002

Taylor, S. F., & Liberzon, I. (2007). Neural correlates of emotion regulation in psychopathology. *Trends in Cognitive Sciences*, *11*(10), 413–418. http://doi.org/10.1016/j.tics.2007.08.006

Tull, M. T., Bardeen, J. R., DiLillo, D., Messman-Moore, T., & Gratz, K. L. (2014). A prospective investigation of emotion dysregulation as a moderator of the relation between posttraumatic stress symptoms and substance use severity. *Journal of Anxiety Disorders*, *29C*, 52–60. http://doi.org/10.1016/j.janxdis.2014.11.003

Tull, M. T., Barrett, H. M., McMillan, E. S., & Roemer, L. (2007). A preliminary investigation

of the relationship between emotion regulation difficulties and posttraumatic stress symptoms. *Behavior Therapy*, *38*(3), 303–313. http://doi.org/10.1016/j.beth.2006.10.001

Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE Transactions on Neural Networks / a Publication of the IEEE Neural Networks Council*, *10*(5), 988–999. http://doi.org/10.1109/72.788640

Vytal, K., & Hamann, S. (2010). Neuroimaging support for discrete neural correlates of basic emotions: a voxel-based meta-analysis. *Journal of Cognitive Neuroscience*, *22*(12), 2864–2885. http://doi.org/10.1162/jocn.2009.21366

Wager, T. D., Phan, K. L., Liberzon, I., & Taylor, S. F. (2003). Valence, gender, and lateralization of functional brain anatomy in emotion: a meta-analysis of findings from neuroimaging. *NeuroImage*, *19*(3), 513–531. http://doi.org/10.1016/S1053-8119(03)00078-8

Weiss, N. H., Tull, M. T., Anestis, M. D., & Gratz, K. L. (2013). The relative and unique contributions of emotion dysregulation and impulsivity to posttraumatic stress disorder among substance dependent inpatients. *Drug and Alcohol Dependence*, *128*(1-2), 45–51. http://doi.org/10.1016/j.drugalcdep.2012.07.017

Werner, K., & Gross, J. J. (2010). Emotion regulation and psychopathology: A conceptual framework. In *Emotion regulation and psychopathology : a transdiagnostic approach to etiology and treatment* (Vol. 22, pp. 211–221). New York, NY : Guilford Press. http://doi.org/10.1016/j.janxdis.2007.02.004

Wirtz, C. M., Radkovsky, A., Ebert, D. D., & Berking, M. (2014). Successful application of adaptive emotion regulation skills predicts the subsequent reduction of depressive symptom severity but neither the reduction of anxiety nor the reduction of general distress during the treatment of major depressive disorder. *PLoS ONE*, *9*(10), e108288.

http://doi.org/10.1371/journal.pone.0108288

Wright, C. I., Martis, B., Schwartz, C. E., Shin, L. M., Fischer H, H., McMullin, K., & Rauch, S. L. (2003). Novelty responses and differential effects of order in the amygdala, substantia innominata, and inferior temporal cortex. *NeuroImage*, *18*(3), 660–669.

Yuan, H., Young, K. D., Phillips, R., Zotev, V., Misaki, M., & Bodurka, J. (2014). Resting-state functional connectivity modulation and sustained changes after real-time functional magnetic resonance imaging neurofeedback training in depression. *Brain Connectivity*, *4*(9), 690–701. http://doi.org/10.1089/brain.2014.0262

## Tables

| SVM Map | Mean Prediction Accuracy | Standard Deviation Prediction Accuracy |
|---|---|---|
| Neu vs. Pos | 74.06% | 10.92% |
| Neu vs. Neg | 87.62% | 9.53% |
| Pos vs. Neg | 76.06% | 13.63% |

*Table 1.* Leave-one-out cross-validation (LOOCV) results for 53-subject group model trained and tested on block mean data.

## Figures

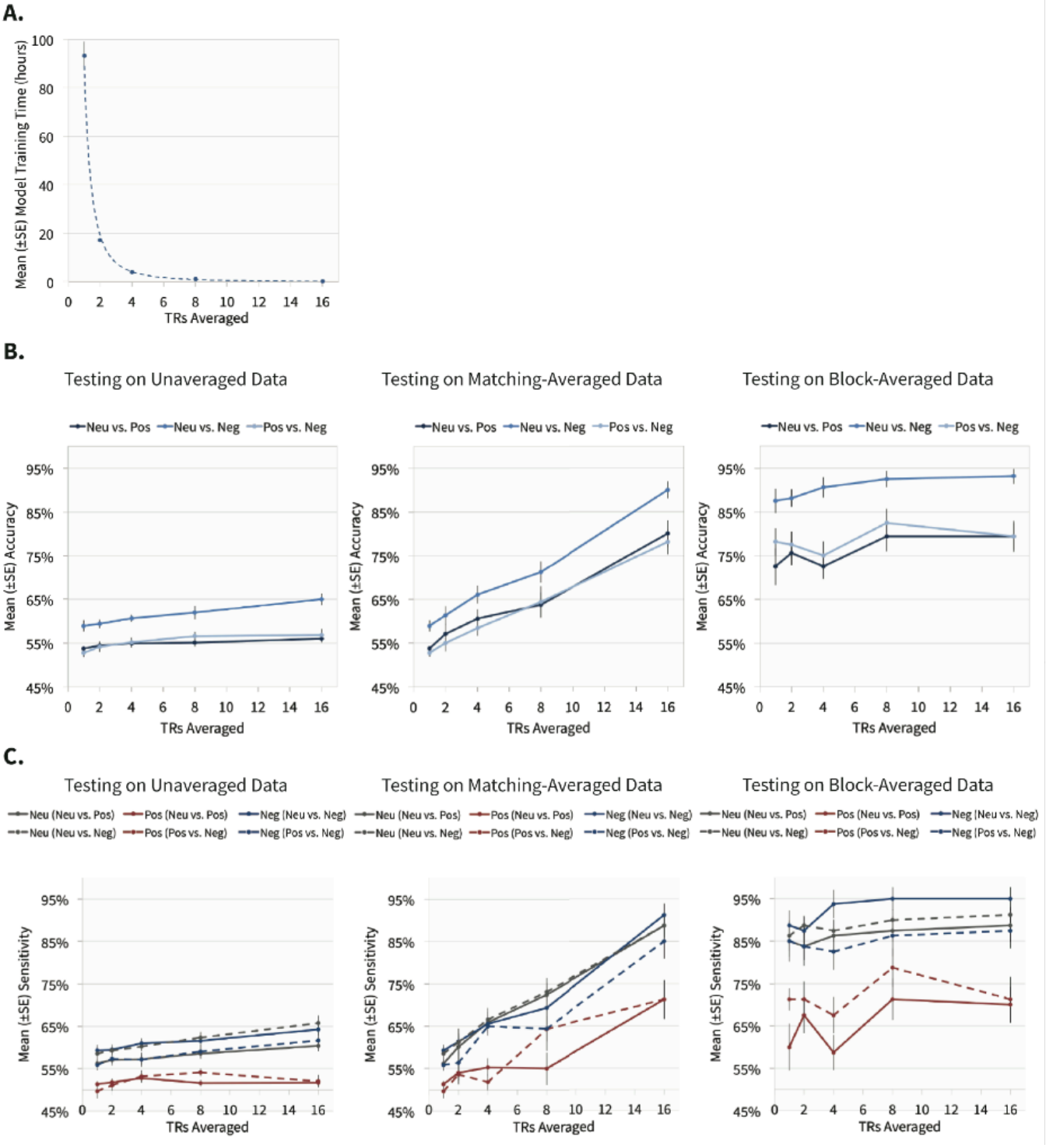


*Figure 1*. Averaging training data across TRs reduces training time and improves model

performance. (A) depicts the effects of averaging training data across multiple time points on the mean (+/- SE) amount of computational time required to train a 16-subject group model using support vector machine learning. (B) and (C) show the effects of averaging training data on (B) mean (+/- SE) prediction accuracy and (C) mean (+/- SE) class sensitivity, respectively. The panels of (B) & (C) illustrate group model performance when testing on data that (left) has not been averaged, (middle) has been averaged across the same number of TRs as the model was trained on, and (right) has been averaged across all TRs in the block.

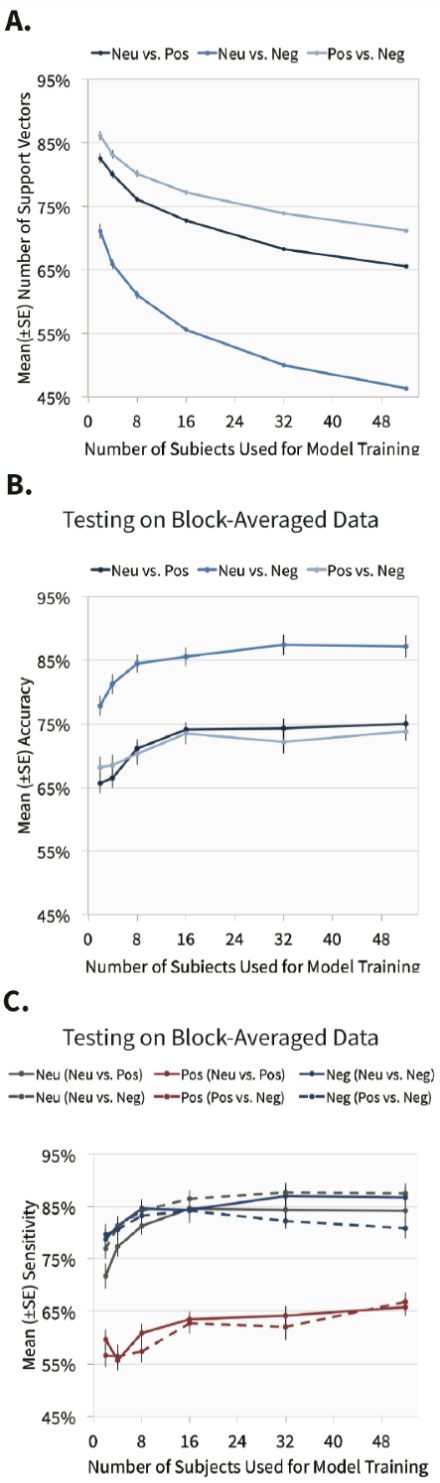


*Figure 2*. Increasing sample size of model training dataset decreases percentage of data identified as support vectors and improves model performance. (A) depicts the effects of varying the

number of subjects used in model training on mean (+/- SE) percentage of training data that are support vectors. (B) and (C) illustrate the effects of varying the number of subjects used in model training on (B) mean (+/- SE) prediction accuracy and (C) mean (+/-SE) class sensitivity, respectively.

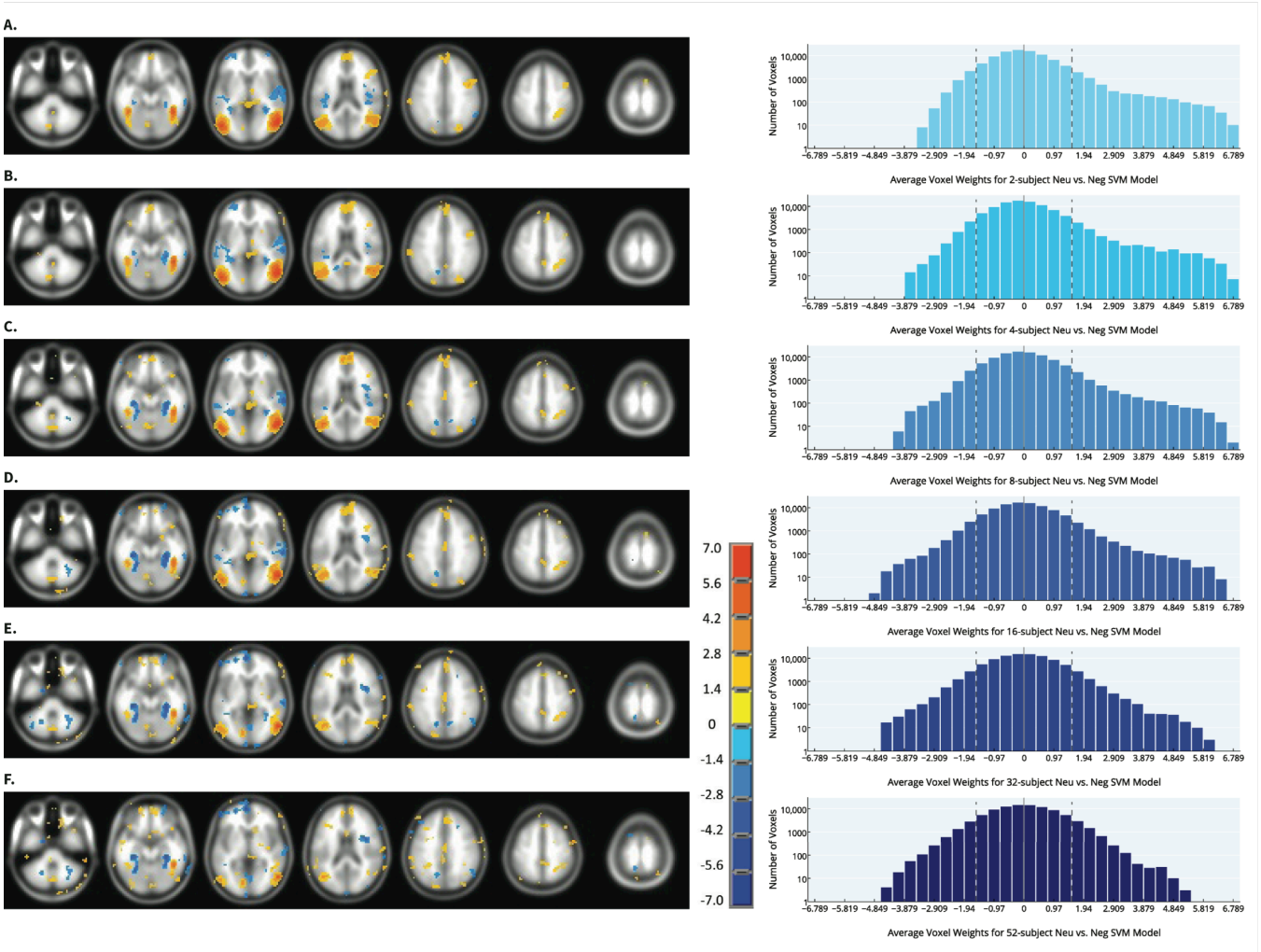


*Figure 3*. Top 10% of mean z-scored weighted voxels for Neu vs. Neg SVM map across group sizes. The left panel displays voxels whose mean weights were in the top 10% of all weights across various group sizes as follows: (A) two-subject model, (B) four-subject model, (C) eight-subject model, (D) 16-subject model, (E) 32-subject model, and (F) 52-subject model. The right panel illustrates the distribution of mean weights for each group size with the threshold for the top 10% indicated with dotted lines. All voxels outside of the dotted lines fall in the top 10%.

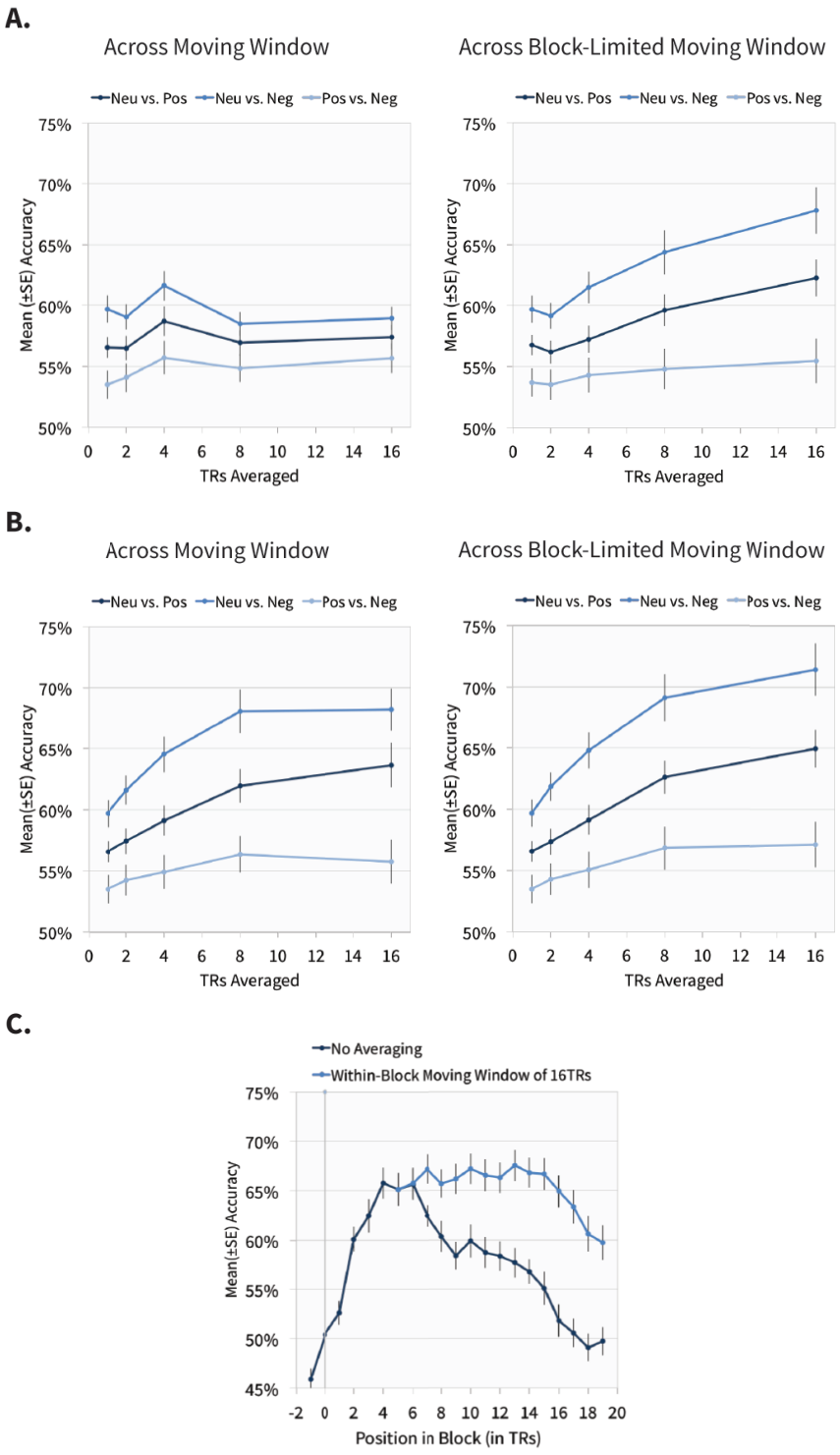


*Figure 4*. Adaptation of group SVM model for real-time fMRI use. (A) and (B) illustrate how temporal compression of testing data (A) across TRs continuously and (B) across TRs only

within blocks impact the mean (+/-SE) accuracy of real-time simulations. (C) and (D) show how temporal compression of the predicted SVM weights (C) across TRs continuously and (D) across TRs only within blocks affect the mean (+/-SE) accuracy of real-time simulations. (E) depicts the mean (+/-SE) accuracy of real-time simulations across the time course of a block for various levels of temporal compression of SVM weights within blocks only.